PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# Computational and experimental investigation of constitutive behavior in AraC

Mary Lowe,[1]* David Gullotti,[1] Ana Damjanovic,[2,3] Ann Cheng,[4] Stephanie Dirla,[4] and Robert Schleif[4]

[1] Physics Department, Loyola University Maryland, Baltimore, Maryland
[2] Department of Biophysics, Johns Hopkins University, Baltimore, Maryland
[3] Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland
[4] Department of Biology, Johns Hopkins University, Baltimore, Maryland

## ABSTRACT

**Many mutations in the N-terminal arm of AraC result in constitutive behavior in which transcription of the *araBAD* genes occurs even in the absence of arabinose. To begin to understand the mechanism underlying this class of mutations, we used molecular dynamics with self-guided Langevin dynamics to simulate (1) wild-type (WT) AraC, (2) known constitutive mutants resulting from alterations in the regulatory arm, particularly alanine and glycine substitutions at residue 8 because P8G is constitutive, whereas P8A behaves like wild type, and (3) selected variant AraC proteins containing alterations in the dimerization core. In all of the constitutive arm mutants, but not the WT protein, residues 37–42, which are located in the core of the dimerization domain, became restructured. This raised the question of whether or not these structural changes are an obligatory component of constitutivity. Using molecular dynamics, we identified alterations in the core that produced a similar restructuring. The corresponding mutants were constructed and their *ara* constitutivity status was determined experimentally. Because the core mutants were not found to be constitutive, we conclude that restructuring of core residues 37–42 does not, itself, lead to constitutivity of AraC. The available data lead to the hypothesis that the interaction of the N-terminal arm with something other than the front lip is the primary determinant of the inducing versus repressing state of AraC.**

Proteins 2014; 82:3385–3396.


**Key words: allostery; gene regulation; self-guided Langevin dynamics; AraC; cluster.**

## INTRODUCTION

Regulation of the L-arabinose operon in *Escherichia coli*, which allows the cells to catabolize arabinose, first became of interest when genetics data suggested that expression might be under positive control of AraC[1,2] rather than the much better characterized (up to that time) negative regulation seen in the *lac* operon and in phage lambda.[3] Subsequent to the demonstration of positive control, the phenomenon of DNA looping, which is required for repression in the *ara* system, was first discovered and demonstrated in the operon.[4–7] In the absence of arabinose, the two DNA binding domains of AraC are rigidly held in positions/orientations that allow the two domains to bind to the $araI_1$ and $araO_2$ DNA sites (Fig. 1), which are separated by 210 base pairs, and disfavor binding to the adjacent direct repeat $araI_1$ and $araI_2$ sites at the *ara* $p_{BAD}$ promoter.[8,9] Upon the binding of arabinose, the DNA binding domains are less rigidly held, and can then more easily bind to the $I_1$ and $I_2$ sites and induce the promoter.

In recent years, much effort has been devoted to determining the molecular mechanism by which the binding of arabinose to the dimerization/arabinose binding domains then allows greater positional and orientational freedom to the DNA binding domains. Key to this transition are the N-terminal 18 amino acids of the protein. These constitute an arm that in the presence of arabinose binds over a bound arabinose molecule,[10] and in the absence of arabinose, are likely differently structured.[11] Most amino acid substitutions in the arm, and largely only mutations in the arm, leave the protein unable to engage in DNA looping,[12,13] and by default, the mutant proteins, called constitutive, activate transcription from

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Mary Lowe; Physics Department, Loyola University Maryland, Baltimore, MD. E-mail: mlowe@loyola.edu

Received 27 January 2014; Revised 20 July 2014; Accepted 31 August 2014
Published online 22 September 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24693

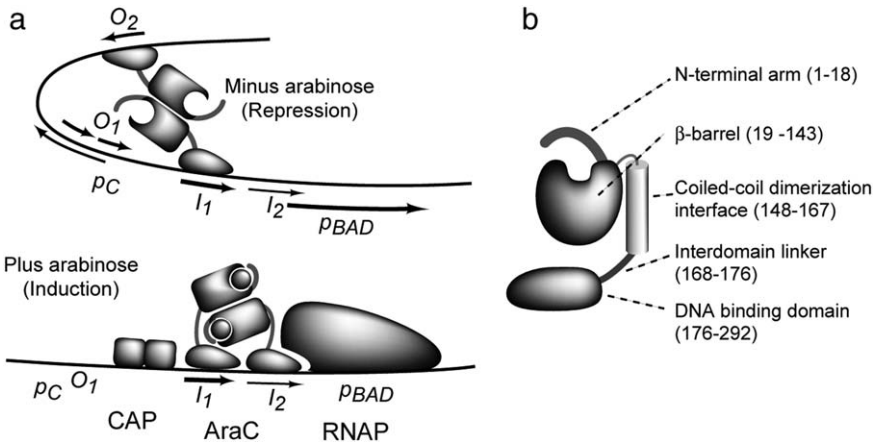


**Figure 1**

The *ara* regulatory region and regulatory proteins. (a) A schematic of the looped repressing and nonlooped inducing states of the arabinose operon. In addition to the $O_2$, $I_1$, and $I_2$ sites that control the activity of $p_{BAD}$, the $O_1$ double site controls activity of the $p_C$ promoter for the synthesis of AraC. The cyclic AMP receptor protein CAP is required for full activity of both promoters. (b) A schematic of one of the subunits of AraC showing its major parts. The β-barrel is the dimerization core.

*ara* $p_{BAD}$. Beyond this, virtually nothing is known about the mechanistic basis of constitutivity. Because arm mutations at only one position of the arm, Phe-15, were instead uninducible, their mechanistic basis has been studied with self-guided Langevin dynamics (SGLD) simulations and found to be based largely on the formation of a hydrophobic cluster whose existence is necessary for the arm to fold properly over a bound arabinose molecule.[14] In the work reported here, we have again used SGLD to examine the mechanistic basis of mutations in the N-terminal arm. In this case, much of the focus was on residue Pro-8, where substitution of a glycine residue yields constitutive behavior, whereas substitution of an alanine residue yields wild-type (WT) behavior.

For this study, we used SGLD, which accelerates systematic low frequency motions of proteins through application of guiding forces determined on-the-fly.[15–17] This method has been used previously to identify conformational changes in several proteins.[14,18–21] The SGLD results were benchmarked against molecular dynamics simulations of these proteins and have been shown to exhibit the same qualitative behavior, whereas the conformational changes in SGLD were observed on a faster timescale.[14,18–21] In particular, in a previous study on AraC,[14] all possible substitutions were simulated at the single position Phe-15 in the N-terminal arm where a noninducible mutation has been found. Notably, the SGLD simulations predicted two new substitutions at the position that would have WT behavior. All the substitutions at the position were then constructed and tested *in vivo*, and all the SGLD predictions were verified experimentally.

In the present work, we simulated a constitutive mutation in the arm, P8G, that induces nearly full expression of the *araBAD* genes in the absence of arabinose, whereas a similar substitution, P8A, leaves the protein with WT behavior.[12] Not only did the P8G mutation change the structure of the N-terminal arm, but compared with a simulation of WT AraC and P8A, we found that the P8G mutation led to an alteration of the structure of a portion of the core of the dimerization domain more than 22 Å away. Understanding the differences among the P8G, P8A, and WT structures should help identify the fundamental cause of constitutivity. SGLD simulation of additional constitutive mutations in the arm, for example, at residues 9, 10, and 11, also displayed similar structural alterations in the core and displayed substantial changes to the structure of the N-terminal arm itself. The fact that P8G and other constitutive mutations in the arm produced the same changes in the core's structure raised the question of whether or not the core change itself is sufficient to cause constitutivity. To address this question, we used SGLD to identify potential mutations in the core that yield predictions of structural changes in the core similar to those observed with P8G. Using molecular genetics, we then constructed these mutant proteins and tested experimentally in the lab whether they were constitutive. As they were not, we conclude that constitutivity does not obligatorily result from the structural change of the core.

## MATERIALS AND METHODS

### Preparation of the initial structures for computation

The program CHARMM[22] and CHARMM27 parameter and topology files were used to prepare the input files for simulations. The parameter file included the

CHARMM22 protein force field, dated August 1999.[23] The starting structure was the apo form of the AraC dimerization domain, Protein Data Bank, entry code 1XJA, chain D, that was crystallized with the Y31V mutation.[11] Residue 31 was restored to tyrosine to achieve the WT sequence. The uncharged form of histidine, HSD, was used for all histidine residues except H80, for which the form, HSE, protonated on NE2, was used because of the proximity of the ND1 atom to Arg-38. Because the sequence in 1XJA started at residue 7 and mutations of residue 8 were being studied, the sequence was extended to residue 6 to reduce end effects. PDB entry 2KHO[24] was found to contain the sequence NDPL corresponding to residues 6–9 of AraC, and its coordinates were used to extend AraC to include residue 6. The final sequence of AraC consisted of residues 6–171. CHARMM was used to add hydrogen atoms, acetylate the N-terminus, and replace the C-terminus with N-methylamide. The system was energy minimized using the steepest descent method for 500 steps.

Crystallographic waters were included, and additional TIP3 waters were added. Water molecules within 2.5 Å of the protein were removed. The protein was centered at the coordinate origin, and all water molecules farther than 36 Å from the origin were removed.[14] Eleven sodium and seven chloride ions were added to neutralize the WT sequence. The system was prepared with rhombic dodecahedral symmetry with a volume equivalent to that of a sphere of water with radius 36 Å. The solvated system was energy minimized using the adopted basis Newton-Raphson method for 180 steps before simulation. AraC mutants P8G, P8A, L9G, L10D, L10A, P11R, R38A, M42A, I46A, and Y97A were prepared similarly. Mutants L10D and R38A were neutralized with 11 sodium and 6 chloride ions; P11R was neutralized with 10 sodium and seven chloride ions.

## Molecular dynamics and SGLD

An ensemble was constructed from a set of molecular dynamics simulations for each mutant in which initial velocities for each atom were determined by a random number generator using a different seed for initial velocities of each simulation. This process resulted in a unique trajectory for each simulation. A time step of 1 fs was used. Rhombic dodecahedral periodic boundary conditions and the particle mesh Ewald method for electrostatic interactions were used, with $\kappa = 0.45$, interpolation order of 6, grid spacing of ~1 Å, and real space interaction cutoff of 10 Å. Heating was conducted from 0 to 300 K over a 100-ps period. Equilibration was performed at 300 K for 500 ps in an NPT ensemble. Constant temperature was maintained using a Hoover thermostat with a coupling constant of 1000 $ps^2$ kcal/mol, and constant pressure was maintained using a barostat with a piston mass of 500 amu and piston collision frequency of 20/ps.[14] Molecular dynamics simulations were performed with SGLD.[15,16] As suggested by previous studies,[18] parameters for SGLD were: guiding factor $\lambda = 1$, local average time $t_L = 0.1$ ps, and collision frequency $\gamma = 1$/ps, which is related to the friction coefficient. The SGLD production runs were performed in an NVT ensemble at 300 K. Coordinates were saved at 1 ps time steps.

For WT and P8G, 10 and 8 simulations were performed, respectively, including heating, equilibration, and 6 ns of SGLD production runs. For the arm mutants, P8A, L9G, L10D, L10A, and P11R, six simulations were performed for 6 ns. Six simulations were performed for each core mutant; R38A was simulated for 6 ns; M42A for 4 ns; I46A and Y97A for 3 ns. This amount of time was sufficient to observe restructuring of residues 37–42.

## Analysis procedures

### Energy matrices

Interaction energies were computed between residues with the INTEraction command in CHARMM. Solvent atoms were not included. Interaction energy matrices were constructed with elements $E_{ij}$ corresponding to the interaction energy between residues $i$ and $j$ for residues 7–171 in a particular frame. The average of the interaction matrices was computed by averaging $E_{ij}$ over frames separated by 20 ps for the last nanosecond of the simulations and over all available seeds. For WT and P8G, 500 and 400 frames were averaged, respectively. Interaction energies between neighboring residues were computed but disregarded because the energies were dominated by covalent bonds.

### Distance matrices

Distance matrices were constructed for residues 6–171 at each picosecond and for all seeds. $D(A)_{ij}$ is an element of a distance matrix between the $C_\alpha$ atoms of residues i and j of protein A, averaged over time and all available seeds. The element $\delta(A,B)_{ij}$ is equal to $D(A)_{ij}$-$D(B)_{ij}$, and displays the internal structural differences between proteins A and B. The matrices δ(P8G, WT) and δ(P8A, WT) were computed. To find the average distances involving residues in regions exhibiting low frequency oscillations, D(WT) was averaged over 6 ns; this is the reference WT distance matrix. In simulations of P8G, it took a few ns for structural changes to occur, and thus, to maximize the structural differences between WT and the mutant, D(P8G) and D(P8A) were constructed for the final ns of simulations only.

### RMSD and fluctuations

For a given mutant, the root mean square deviation, RMSD, was calculated at each time point for each residue of a seed using backbone atoms C, N, $C_\alpha$, and O.

The energy minimized structure for that mutant was the reference structure. The following formula was used:

$$RMSD_{ijk} = \sqrt{\frac{d^2_{Nijk} + d^2_{C\alpha ijk} + d^2_{Cijk} + d^2_{Oijk}}{4}}$$

where *i*, *j*, and *k* are the seed number, the frame number indicating the time point, and residue number, respectively, and *d* is the distance between equivalent atoms of the evolved structure and the reference structure. The average RMSD is the average over seeds, time, and residues:

$$Average\ RMSD = \frac{\sum_{k=init}^{final} \sum_{j=1}^{N} \sum_{i=1}^{N_s} RMSD_{ijk}}{(final - init + 1)\, N \cdot N_s}$$

where $N$ is the number of frames, $N_s$ is the number of seeds, *init* and *final* are the first and last residues in the region of interest. For example, the lip region corresponds to *init* = 37 and *final* = 42.

For each simulation, RMSD values were calculated for frames recorded every 10 ps. Three methods of displaying the values were implemented. (1) To compare simulations for different mutants, average RMSD values were computed for each residue over the final ns of simulation time and all available seeds. (2) To look at the arm and lip regions only, the average RMSD was calculated over arm residues 7–18 and lip residues 37–42 during the last ns for all seeds. (3) Average RMSD values for the arm and lip were computed for each frame of each seed. Histograms of the arm RMSDs and lip RMSDs for all seeds were constructed during the last ns. For example, the WT histogram consists of 1000 values, that is, RMSD values from 100 frames for each of 10 seeds.

### Fitting an ellipsoid and calculation of an angle structure parameter

To determine the orientation of residues 37–42, an ellipsoid was fitted to the backbone atoms of 37–42 by calculating a covariance matrix (second central moment matrix) for the coordinates of the position vectors of the atoms. Programs for fitting and vector operations were written in Mathematica. In all cases, the smallest eigenvalue was a factor of 5–10 times smaller than the largest eigenvalue and 3–5 times smaller than the middle eigenvalue, indicating that one of the three axes of the ellipsoid is much shorter than the other two; that is, the atoms lie close to a plane perpendicular to the short axis of the ellipsoid. This plane is normal to the eigenvector associated with the smallest eigenvalue.

A structure parameter $\theta$ was computed. A vector **V** was constructed between the center of points of the backbone atoms of residues 37 and 42 and the center of points of the backbone atoms of the top β-strand in the back of the core (residues 19–26). The β-strands in the back of the core do not exhibit a structural difference between WT and P8G. At the last time point in each simulation, the angle between **V** and the normal to the plane spanning residues 37–42 was calculated. This was repeated for all seeds of the mutant at the last time point of the simulations.

### Correlation analysis

Correlated movements of $C_\alpha$ atoms were examined using Pearson cross correlation coefficients.[19] The matrix elements $C_{ij}$ were calculated as follows over a particular time interval of the simulations:

$$C_{ij} = \frac{\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \rangle}{\sqrt{\langle \Delta\vec{r}_i(t)^2 \rangle \langle \Delta\vec{r}_j(t)^2 \rangle}}$$

The structures in all frames were initially aligned with the structure in the first frame. The time-averaged structure was computed. The vector $\Delta\vec{r}_i(t)$ is the displacement of the position vector $\vec{r}_i(t)$ from its time-averaged position. The angle brackets $\langle\rangle$ refer to a time average. The $C_\alpha$ of all residues 6–171 were included in the $C_{ij}$ matrix. Correlation coefficients were determined for individual simulations and were further averaged over all seeds. Alignments were accomplished in VMD, and correlations were computed in MATLAB.

### Collision analysis

Python scripts were used to analyze the simulation files spanning 6 ns of production. For each time point of a simulation, interatomic distances were calculated between each atom of a selected residue and all other atoms in the protein. Each distance was compared with the sum R of the van der Waal's radii of the pair of atoms. If the distance was less than αR, where α is a constant ~1, a collision was tallied. Frequent collisions between pairs of atoms were noted. Collisions between atoms of adjacent residues were disregarded because they occur at each time step and were not informative for distinguishing between WT and mutant behavior. Collisions were useful for identifying clusters of side chains.

### Experimental methods

The WT *araC* gene was previously cloned between the *NcoI* and *SacI* sites in the multiple cloning region of the pET24d vector.[25] Mutations were introduced into *araC* using QuikChange Site-Directed Mutagenesis (Stratagene). Plasmid DNA was isolated from single colonies of DH5α cells and sequenced to confirm the mutations.

Plasmid DNA containing WT or mutant *araC* was transformed into SH321[26] for measurement of the ability of AraC to induce *ara* $p_{BAD}$ and into SH10 (referred

to as DMH90[27]), which contains a *pC-lacZ* fusion for measurement of the ability to repress *ara* $p_C$ by measurement of β-galactosidase.[28] Arabinose isomerase assays to determine uninducibility and constitutivity were performed on cells growing exponentially in M10 minimal medium[29] supplemented with 40 μg/mL kanamycin, 0.4% (v/v) glycerol, 10 μg/mL thiamine, 20 μg/mL L-leucine, 1% (w/v) casamino acids, and 0.2% (w/v) L-arabinose, when present. Because of daily fluctuations in the isomerase assay, for comparison to previous experimental determinations, enzyme levels were compared with the level of arabinose isomerase in the same cells but containing WT AraC growing in the same medium at the same time.

To determine experimentally whether candidate mutations in the core of AraC were folded and possessed the ability to dimerize and bind to DNA to repress, we used the fact that WT AraC in the repressing state represses activity of the *ara* $p_c$ promoter and that this activity can be easily determined by assaying for β-galactosidase[28] in a $p_c$-*lacZ* fusion.

## RESULTS

SGLD was used to accelerate systematic, low frequency motions of the protein, to sample a larger conformational space.[15,16] The guiding force in SGLD accelerates the motion so that 1 ns of simulation does not correspond to 1 ns of protein motion.[15] SGLD has been benchmarked against MD simulations in its ability to sample protein conformational rearrangements in several proteins, including AraC as described in the introduction. In all cases, SGLD showed the same structural trends as MD, but the magnitude of observed conformational changes was larger.[17–20] In this article, simulations using SGLD were conducted for WT, P8G, P8A and other mutants, where the initial structure was the WT structure modified by the appropriate sequence mutations. Multiple trajectories of each mutant were simulated using different seed values.

### Structural differences among P8G, P8A, and WT

This work was centered on residue 8 because at this position a substitution to glycine yields constitutive behavior of AraC, whereas an alanine substitution yields WT behavior. Hence, we sought to identify structural differences among the P8G, P8A, and WT proteins.

#### *RMSD values*

The average RMSD during the sixth ns of all simulations is shown in Figure 2(a). The most prominent differences among WT, P8A, and P8G occur at residues 37–42 and the region around 31. Because the latter is a highly mobile loop far from the arm, we therefore considered the changes in the region 37–42 to be more significant. WT and P8A behave similarly, whereas P8G is different, which is consistent with experiments. Residues 37–42 lie in the dimerization core (Fig. 1).

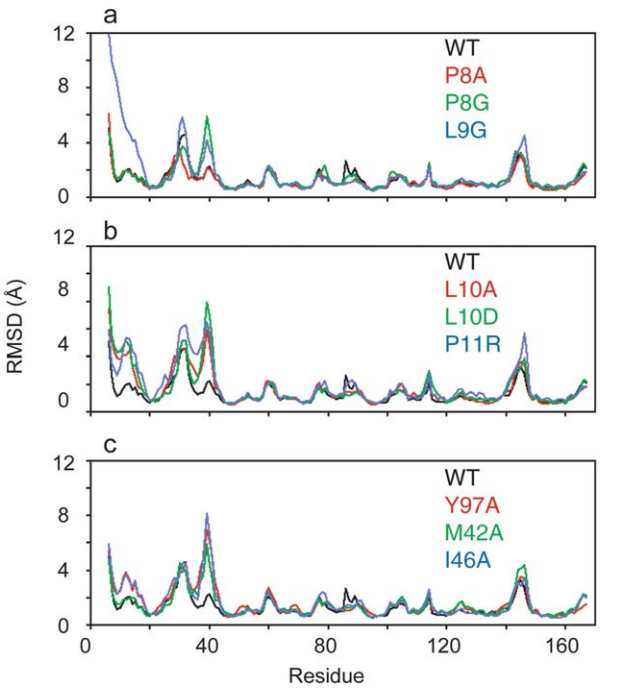


**Figure 2**

RMSD values in Å for WT and mutants as a function of residue number in the last nanosecond of the simulations. (a) Average RMSD for each residue for WT (black), P8A (red), P8G (green), and L9G (blue). (b) Average RMSD for WT (black), L10A (red), L10G (green), and P11R (blue). (c) Average RMSD for WT (black) and core mutants Y97A (red), M42A (green), and I46A (blue). The average RMSD was computed over time and all simulations (seeds) of a mutant.

#### *Distance difference matrices*

To understand the structural changes of residues 37–42 in more detail, we used averaged distance matrices and distance difference matrices (see Methods). To verify that the ensemble of structures sampled during a production run was in steady state, we checked that the distance matrices for WT over two nonoverlapping intervals of the production run were highly similar. A distance matrix was constructed by averaging values at each picosecond from 0 to 3 ns for 10 simulations. To ascertain that the simulations for WT had converged, a similar matrix was constructed for 3–6 ns. The two matrices were highly similar (Supporting Information Fig. S1), indicating that WT is stable. Averaging reduced the effects of thermal noise and most low frequency oscillations, such as those found in loops, but there were still small structural differences involving residues 31–32 and 141, which lie at the ends of β-strands or α-helices.

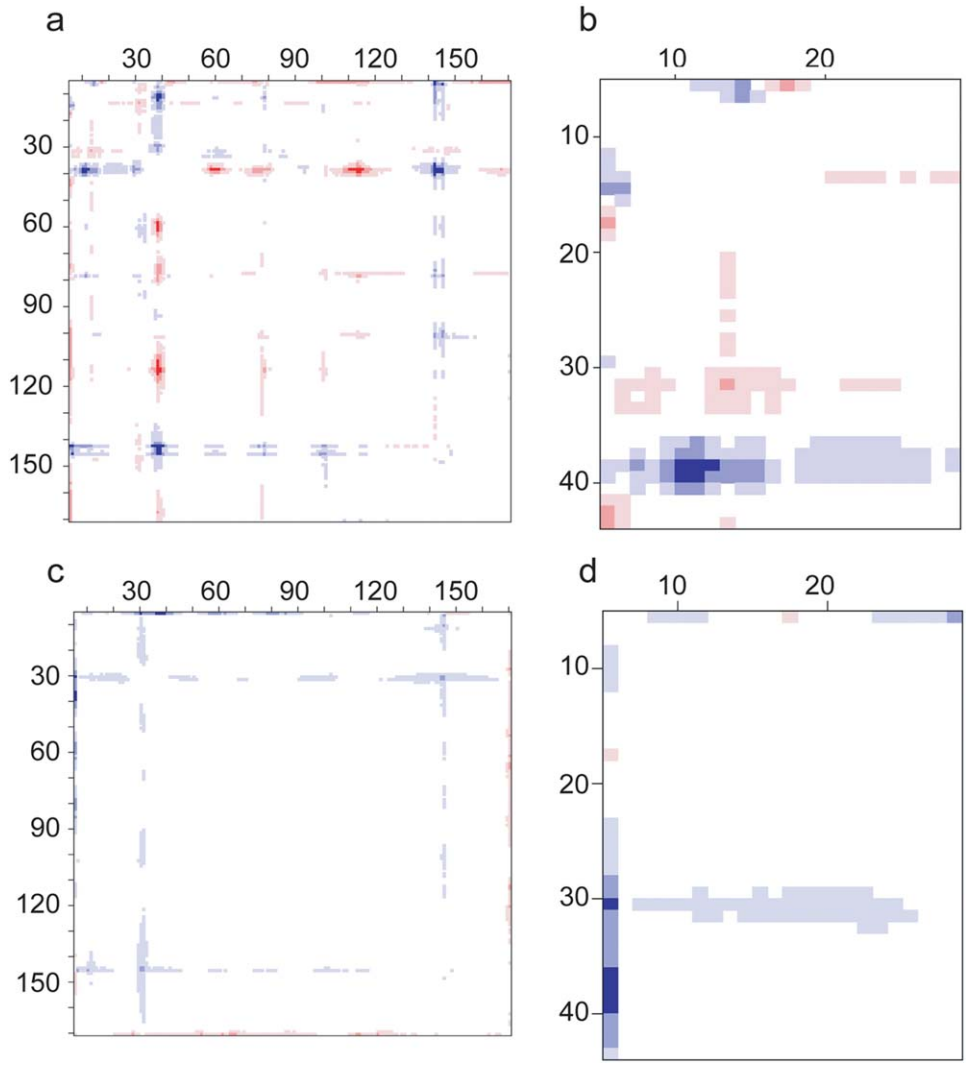


**Figure 3**

The difference of distance matrices between P8G and WT and between P8A and WT. (a) The difference δ(P8G, WT) is shown for residues 6–171. Residue numbers are indicated on the left and top. P8G is constitutive. (b) Expanded view of the upper left corner of δ(P8G, WT) showing how residues 37–41 change position relative to the arm. (c) Distance difference matrix δ(P8A, WT). P8A behaves like WT. (d) Expanded view of the upper left corner of δ(P8A, WT). Green corresponds to a difference in distances from $\leq -4$ Å, dark blue ($-4$, $-3$), medium blue ($-3$, $-2$), light blue ($-2$, $-1$), white ($-1$, 1), light red (1, 2), medium red (2, 3), dark red (3, 4), and green $\geq 4$ Å.

To determine the structural difference between P8G and WT, the distance difference matrix δ(P8G, WT) was computed using the reference WT distance matrix and the distance matrix D(P8G) (see Methods and Supporting Information Fig. S2). As shown in Figure 3(a,b), notable differences between WT and P8G involve residues 6, 14, 31–34, 37–41, 78, 101–102, and 143–146. Residues 31–34, 78, 101–102, and 143–146 correspond to flexible loops. Residue 6 is at the tip of the arm and is flexible. Thus, the regions most likely containing significant structural differences between P8G and WT are near residues 14 and 37–41.

The locations of major features of AraC are shown in Figures 1 and 4. In Figure 3(a,b), the difference between the positions of 37–41 and arm residues 6–17 is depicted in blue, indicating a decrease in the distance relative to WT. There is also a decrease in distance between 37–41 and the loop 143–146 between the two α-helices. The largest distance difference of $-4.7$ Å occurs between residues 39 and 143. The difference between 37–41 and most of the rest of P8G is depicted in red, indicating an increase in distance. Together the results of the distance matrix and RMSD analyses indicate that core residues 37–42 are most easily restructured in P8G.

For comparison, the distance difference matrix δ(P8A, WT) is shown in Figure 3(c,d). Most of the matrix does not indicate a significant difference in structure between P8A and WT. There are differences near residues 31 and 146, but these regions correspond to flexible loops.

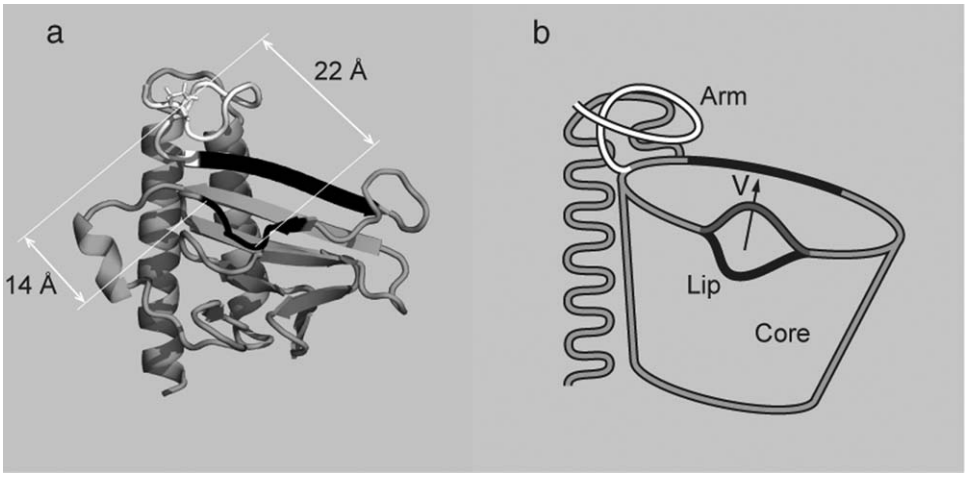


**Figure 4**

Representations of WT AraC. (a) The arm is shown in white. Residues 37–42 (black) are located at the front lip. Pro-8 (white stick representation) is located in the arm. The top back β-strand is shown in black. The distance between $C_\alpha$ of Pro-8 and $C_\alpha$ of Pro-39 is 22 Å. The distance between $C_\alpha$ of Pro-8 and $C_\alpha$ of Met-42 is 14 Å. (b) Cartoon. The dimerization core/arabinose-binding domain contains a "cup" and α-helices. The ligand arabinose sits inside the cup near the lip. The backbone of residues 37–42, in the WT core, is largely in the plane of the top of the cup, black. In constitutive mutants, this region is flipped upwards as shown in dark gray. The vector **V** extends across the top of the cup from the center of the lip to the center of the back β-strand.

### *Correlation analysis*

Clusters of residues that tend to move together can sometimes be identified by Pearson correlation coefficients (see Methods). Coefficients $C_{ij}$ were computed over 6 ns and all seeds for WT and P8G, and the difference in the matrices, C(P8G) − C(WT), was constructed (Supporting Information Fig. S3). The correlation matrices showed that residues 7 and 11 are more positively correlated in WT ($C_{7,11} = 0.28$) than in P8G ($C_{7,11} = 0.03$),. In addition, there are weak differences Δ in $C_{ij}$ between the arm and the β-strands at the back of the cup, for example, between residues 12–14 and 20–22 ($\Delta = 0.11 \pm 0.01$) and between 8–14 and 45–53 ($\Delta = 0.12 \pm 0.03$).

### *Interaction energies*

In Figure 5, we constructed energy matrices containing the complete set of inter-residue interaction strengths over the entire protein. Let matrix element $E(A)_{ij}$ be the ensemble average of the energy interaction between residues *i* and *j* for protein A. Figure 5 shows the energy difference matrix, E(P8G) - E(WT). Residue pairs along the diagonal and in the elements next to the diagonal, which reflect covalent bonding, are ignored.

Differences greater than 0.3 kcal/mol between WT and P8G appear throughout the protein. Although thermal energies at 300 K are 0.6 kcal/mol, we consider energy differences greater than 0.3 kcal/mol to be worth noting because of the extensive averaging of the energy matrices. A high density of significant energy differences may be seen within the arm region (residues 7–18) indicating that the region as a whole changes its interaction within itself, confirming structural changes in the arm region. Energy interaction differences occur between Arg-38 and residues 9, 11. No other interaction differences are observed between the arm and residues 37–42. We note that Arg-38 is located in the lip region formed by residues 37–42, a region that becomes restructured in P8G and comes closer to the arm, as indicated by the distance matrix (Fig. 3). The fact that Arg-38 exhibits the most substantial change in the interaction energy is due to the fact that it is the

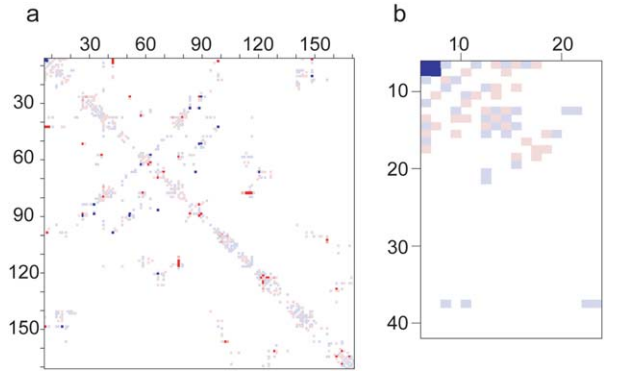


**Figure 5**

Interaction energy difference matrix E(P8G) – E(WT). (a) Matrix for residues 7–171. The energy matrices for WT and P8G represent the ensemble average over 500 frames and 400 frames, respectively, during the last nanosecond. Residue numbers are indicated on the left and top. Color code with values in kcal/mol: ≤ −3 dark blue; (−3, −2) medium blue; (−2, −0.3) light blue; (−0.3, 0.3) white; (0.3, 2) light red; (2, 3) medium red; (3, 100) dark red. (b) Expanded view showing the interactions within the arm residues 7–18 and between the arm and Arg-38.

only charged residue from that region. The interaction energy involving Arg-38 is likely exaggerated in our energy matrix, because the calculations are not taking into account the screening by water molecules.

Based on Figure 5, the strengths of interaction energies are also high between the arm and the regions surrounding residues 95–101 and 141–149, and large differences in these interaction energies exist between P8G and WT. Although in this article we focus on residues 37–42, these other regions could be the basis of a future study.

### Characterization of structural changes through angle $\theta$

To characterize the structural changes of residues 37–42, we sought a structure parameter that would describe the conformational change more precisely. Visual observation and eigenvalues of the covariance matrix reveal that in WT, the backbone of residues 37–42 is well approximated by a U-shape that lies in a plane. The angle $\theta$ between the normal to the plane and the vector **V** (see Fig. 4 and Methods) was $87^\circ \pm 28^\circ$ for WT and $66^\circ \pm 37^\circ$ for P8G. Relative to WT, the U-shaped region in P8G flipped up toward the arm leading to restructuring of residues 37–42, which we call the "lip." The large standard deviation of the P8G angle (Table I) is due to the fact that not all seeds exhibit restructuring. For WT, the large standard deviation is due to one outlier angle.

A comparison of $\theta$ with the average RMSD for lip residues 37–42 shows that larger values of the average RMSD are correlated with smaller values of $\theta$ (Table I). WT and P8A have the smallest average RMSD values of the lip. To understand why the standard deviations in $\theta$ are large, histograms of the lip RMSDs during the last ns are shown in Figure 6. The lip distributions for WT and P8A are similar and have small standard deviations, whereas for P8G, the distribution is much broader.

### Structural changes through regular molecular dynamics simulations

As a result of the guiding force that is applied in SGLD to accelerate motion, a question arises whether SGLD may magnify motions in AraC beyond what might be seen using regular molecular dynamics. To check the presence of artifacts in the SGLD results reported here, we performed five independent, conventional MD runs of the mutant P8G spanning 25 ns. In two of the five runs, we observed the same destructuring of residues 37–42 as we observed with SGLD. This result is consistent with the wide range of angles observed in SGLD simulations of P8G (Table I) and in the broad RMSD distributions of the lip shown in Figure 6. We do not see evidence that SGLD is generating artifactual conformational changes in the AraC system. We also note that in a previous investigation comparing MD and SGLD using the same, relatively small guiding forces used in this work, SGLD was found to be fully consistent with MD in representing structures and structural relaxation, while at the same time generating these structural changes on a faster timescale.[14,18,19,21]

## Identification of core mutants producing restructuring of residues 37–42

The structural difference in the core between P8G and either WT or P8A raises several questions. Measurements using DNA as a tape measure suggest that in the repressing state of AraC, the DNA binding region (DBD) may be in contact with the front of the core.[30] Therefore, restructuring of residues 37–42, which lie in the front of the core, Figure 4, might cause the DBD to be released in the absence of arabinose, and thus lead to expression, that is, constitutivity. The question arises as to whether it is possible to change the conformation of this region by mutating the core. If molecular dynamics indicated such a structural change, it would then be possible to construct the mutant and determine experimentally whether the mutant protein is constitutive. A second question raised by the differences between P8G and WT concerns the distance, 22 Å, between the mutation and the structural change. How is it possible for the effects of a mutation at residue 8 to be transmitted to such a distant location? To address the latter question, we tried first to understand with additional analysis of SGLD simulations what caused the restructuring of residues 37–42. Then, we used the information to develop a model for finding mutants of the core that produce a similar restructuring of residues 37–42.

We turned to collision analysis (Supporting Information Fig. S4), visual inspection, and geometric analyses to find that the side chains of Leu-9, Leu-10, Arg-38, Met-42, Ile-46, and Tyr-97 are in close proximity to each other and form a cluster of side chains linking the arm and the front of the core. Figure 7 shows the location of these residues. Residue 8 is not directly part of this cluster, but mutating residue 8 alters the structure of the arm, which could alter the side chain cluster between the arm and the core, and thereby affect the backbone of residues 37–42.

Because collision analysis and visual inspection identified residues 9 and 10 to be clustered with residues 38 and 42, we hypothesized that mutations at or near these arm residues would perturb the interactions between the arm and the core. We then performed SGLD simulations on the arm mutants. Indeed, in these simulations, we found that constitutive mutants L9G, L10D, L10A, and P11R also generated structural changes in the front lip similar to those found for P8G: $\theta$ was small indicating the U flipped up (Table I). On the other hand, for the nonconstitutive mutant P8A, our simulations showed that $\theta$ remained large, similar to WT, indicating no change in the orientation of the U.

Core mutants, R38A, M42A, I46A, and Y97A, were subjected to SGLD simulations and were found to produce a similar restructuring at residues 37–42 as the constitutive arm mutants. M42A, I46A, and Y97A exhibited

**Table I**
Angle of the Lip, Average RMSD Values, and Experimental Classification of Expression

| Protein | Average angle θ (degrees)[a] | Average RMSD of arm, Å[b] | Average RMSD of lip, Å[b] | Phenotype classification |
|---|---|---|---|---|
| WT | 87 ± 28 | 1.6 ± 0.6 | 1.7 ± 1.0 | Inducible |
| P8A (arm) | 95 ± 18 | 1.6 ± 0.8 | 1.7 ± 0.6 | Inducible |
| P8G (arm) | 66 ± 37 | 1.7 ± 0.4 | 3.7 ± 2.1 | Constitutive |
| L9G (arm) | 76 ± 24 | 5.5 ± 3.1 | 2.9 ± 1.5 | Constitutive |
| L10D (arm) | 54 ± 26 | 3.8 ± 1.2 | 5.1 ± 2.0 | Constitutive |
| L10A (arm) | 63 ± 33 | 3.4 ± 2.5 | 3.7 ± 2.1 | Constitutive |
| P11R (arm) | 65 ± 32 | 3.7 ± 0.9 | 4.7 ± 2.3 | Constitutive |
| M42A (core) | 64 ± 18 | 1.7 ± 0.7 | 3.7 ± 1.7 | Uninducible |
| I46A (core) | 46 ± 19 | 2.8 ± 1.1 | 5.0 ± 1.4 | Uninducible |
| Y97A (core) | 71 ± 26 | 2.7 ± 0.8 | 5.0 ± 2.6 | Uninducible |
| R38A (core) | Not measurable | Not measured | Not measurable | Uninducible |

[a]The angle $\theta$ was determined between the vector **V** across the top of the cup and the normal to the plane spanning the backbone atoms of residues 37–42. The average $\theta$ was calculated at the last time point in the simulations for all of the seeds. The uncertainty is the standard deviation of the $\theta$ measurements of the seeds. R38A exhibits extensive repositioning and conformational changes of the front lip, which cannot be measured by the technique used for the other mutants.
[b]The average RMSD and standard deviation were computed for all seeds and all frames during the last ns of the simulations.

a small $\theta$ and large average RMSD of the lip (Table I). R38A resulted in a different shape and repositioning of residues 37–42 and surrounding regions, which cannot be described simply by $\theta$. The average RMSD of the core mutants, Figure 2(c), is similar to that of the arm mutants, Figure 2(a,b). Like the constitutive arm mutants, the core mutants exhibit considerable variability in the RMSD values for the arm and lip (Fig. 6). These core mutants were then tested experimentally for their levels of constitutivity.

## Constitutivity is not due to a change in the core

The preceding analysis found that five different single mutations in the N-terminal arm of AraC, which possess

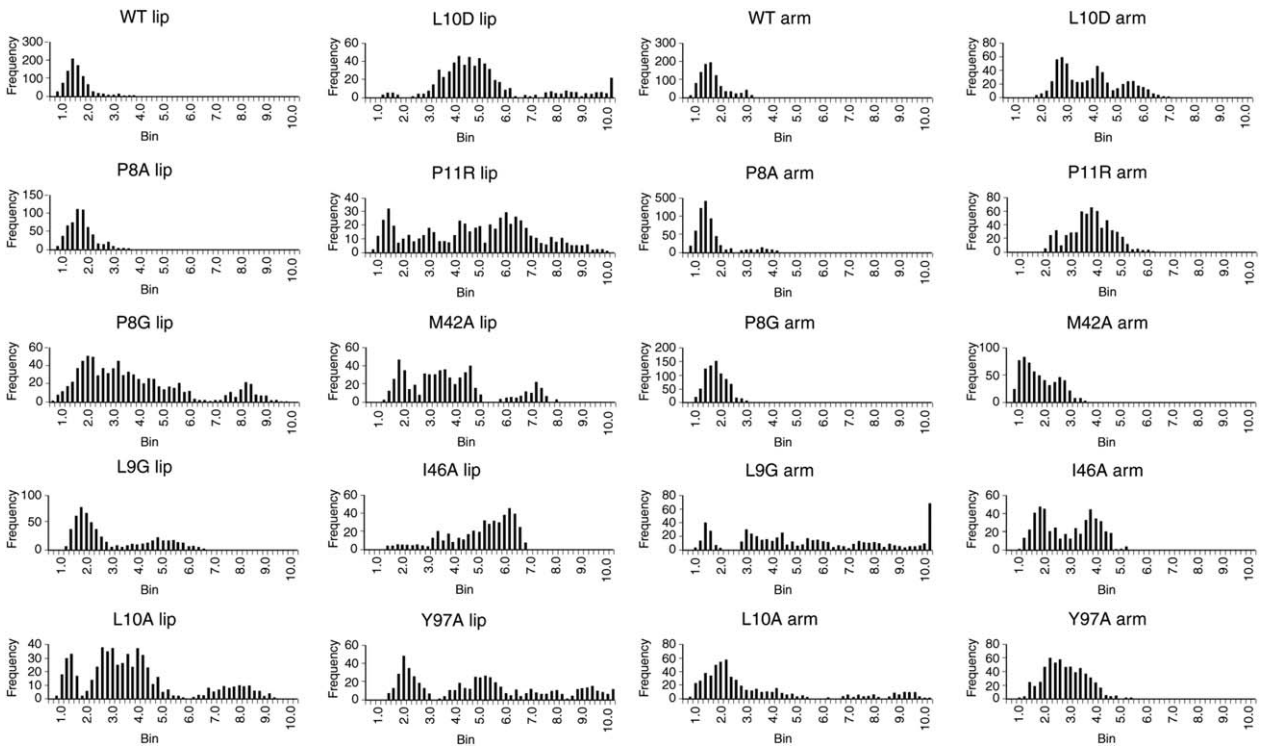


**Figure 6**
Histograms of the RMSD values of the arm and lip residues for WT and mutants. Average RMSD values were computed for each seed in each frame separated by 10 ps for arm residues 7–18 and lip residues 37–42. One thousand values are tabulated for WT; 600 values are tabulated for all mutants except I46A, which has 500 values.

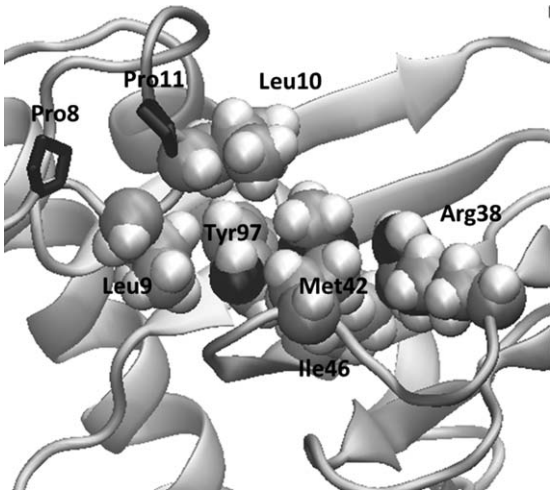


**Figure 7**

Arm and core clusters in AraC. Pro-8 is not part of the arm cluster, but mutations at Pro-8 cause shifts in the cluster. All residues are shown in Van der Waal's representation except for Pro-8 and Pro-11, which are shown in stick form.

constitutive behavior, all led to significant structural changes of residues 37–42 in the front lip. In contrast, WT and one nonconstitutive mutant in the arm did not lead to structural changes of residues 37–42. These results suggest that restructuring of residues 37–42 may be required for constitutivity. To test this possibility, we identified a cluster linking the arm and core and used SGLD to identify mutants in the core of AraC that led to restructuring of residues 37–42; these mutants were M42A, I46A, Y97A, and R38A. We then constructed and measured experimentally the regulatory properties of the mutants. Experimentally, mutants were screened for both their inducing and repressing abilities (Supporting Information Table SI). These core mutants were not found to be constitutive in laboratory experiments.

Because mutants R38A, M42A, I46A, and Y97A were uninducible as well as not constitutive, it is possible that *in vivo*, the proteins did not fold and were therefore totally inactive. We tested this possibility by showing that all of the mutant proteins except Y97A appeared to fold normally because they possess their normal repressing activity in the absence of arabinose (Supporting Information Table SI). This was assayed using an *ara* $p_C$-*lacZ* fusion. WT AraC in the absence of arabinose represses activity of the *pC* promoter so that only about 20 units of β-galactosidase are synthesized in a $p_C$-*lacZ* fusion, whereas in the absence of folded AraC, the level is around 200 units.

Overall, then we find that restructuring of residues 37–42 per se, as predicted by SGLD, is not responsible

**Table II**

Properties of Mutants in the Arm and Core[a]

| | P8 | L9 | L10 | P11 | G12 | Y13 | S14 | R38 | M42 | I46 | Y97 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Trp | | | | | | | | | | | |
| Tyr | | | | C | | *** | C | | | | *** |
| Phe | C | | | C | **I, S** | | | | | | |
| Met | | I | C | | | | | | *** | | |
| Ile | C | | I | | | | | | | *** | |
| Leu | I | *** | *** | | C | | I | **U** | | | |
| Val | **I, S** | C | | | | | I | | | | |
| Thr | | | C | C | C | C | I | | | | |
| Pro | *** | C | C | *** | | C | C | | | | |
| Cys | I | C | I | C | I | U | C | | | | |
| Ser | I | C | I | C | C | C | *** | | | | |
| Ala | **I, N** | C | **C, S** | | I | | C | **U, S** | **U, S** | **U, S** | **U, S** |
| Gly | **C, S** | **C, S** | | | *** | I | | **U** | | **U** | |
| Arg | C | | C | **C, S** | C | C | I | *** | | | |
| Lys | | C | | | | C | | | | | |
| His | C | C | | | | C | | **U** | | | **U** |
| Glu | | | C | | | C | C | | | | |
| Gln | | | C | | | | | **U** | | | |
| Asn | | | C | C | | | I | | | | |
| Asp | | C | **C, S** | C | C | | | | | | |

[a]Mutants were classified as inducible (I) if experimentally, they induce greater than 20% of wild type. Mutants were constitutive (C) if they expressed experimentally at least 20% of the wild-type induced level in the absence of arabinose, uninducible (U) if induction was experimentally less than 20% of wild type, (S) if a structural change in the front lip was present in the MD-SGLD simulations, and (N) if restructuring was not present computationally. Results in bold are reported for the first time in this article and the others are from Ross et al.[12]

for the constitutive regulatory behavior of AraC. As most mutations in the N-terminal arm of AraC do lead to constitutivity (Table II), the data suggest that the arm primarily controls inducibility, independent of the structure of residues 37–42.

## DISCUSSION

The initial goal of this study was to see whether molecular dynamics simulations could provide an insight into the mechanism of constitutivity and a mechanistic explanation for the fact that the P8G mutation makes AraC constitutive and express the *araBAD* genes in the absence of arabinose, the normal inducer of their expression. Simulations performed with SGLD revealed structural changes in the N-terminal arm that contains the P8G mutation and significant structural changes in the core of the dimerization domain, particularly in the lip region at residues 37–42. We found that other arm mutants that are constitutive also produced structural changes in the arm and lip and that a nonconstitutive mutant of the arm, P8A, did not lead to structural changes in the lip (Table I and Figs. 2 and 3). These results raised the question of whether constitutivity is the result of changes in the arm, changes in the lip, or changes in both. To test this idea, we identified residues in the dimerization core, Arg-38, Met-42, Ile-46, and Tyr-97, that are part of a side chain cluster, including arm residues Leu-9 and Leu-10. Indeed, simulations revealed that changing these core residues to alanine resulted in expected restructuring of the lip. Experimentally, however, we found that *in vivo* these mutants are not constitutive. One additional test had to be performed, however. Because they also are not inducible, the possibility existed that they did not fold *in vivo*. We showed that except for Tyr-97, they do fold *in vivo*. We therefore conclude that restructuring of the lip of the dimerization core is insufficient to generate constitutivity in AraC.

If the structure of the lip of the core is not the major determinant of constitutivity of AraC, what is? Because almost all mutations in the arm of *araC*[12] and deletions removing parts of the arm[31] lead to constitutivity, and because the arm repositions itself upon the binding of arabinose,[10,11] the arm must play an important role in communicating the inducing–repressing status of the dimerization domain to the DNA binding domains, but what and how this is accomplished remains largely unknown. Based on RMSD, the backbones of the arms for WT, P8G, and P8A are similar, Figure 2(a). However, the energy interaction matrix, Figure 5, shows differences in the arm between WT and P8G, suggesting that the orientation of the side chains alters and that the coordinates of the polypeptide backbone and their fluctuations should not be used exclusively for understanding the characteristics of the arm. In addition, the distance difference matrix, Figure 3, shows that $C_\alpha$ of Ser-14 increases its distance from many regions of the dimerization core suggesting that the arm as a whole may be shifted slightly in P8G. The distance matrix for the nonconstitutive mutant P8A does not show any change in distance involving Ser-14. Cross correlations show that making a P8G mutation decreases the positive correlation between residues 7 and 11. Taken all together, the P8G mutation alters the structure and/or dynamics of the arm, probably because of the increased flexibility allowed by the proline to glycine substitution. This, in turn, alters a cluster of residues linking the arm and core, and finally, that the core of the dimerization domain changes, but as shown here, the most prominent of these changes, the destructuring of the front lip of the core, does not seem to be required for producing the constitutive response. The P11R mutation results in a shift in Leu-10, which also affects the cluster, and also leads to a structural change in the front lip.

Small hydrophobic clusters that are important for the functional behavior of the protein have been noted before. An intramolecular mechanism was proposed for MAP kinase ERK2 where a hydrophobic cluster consisting of four residues, including a "gatekeeper," was identified.[32] There, information is transmitted through hydrophobic side chain interactions to a network of residues in a different structural motif within the protein. In work on AraC in the presence of arabinose, molecular dynamics simulations showed that residues Leu-9, Tyr-13, Phe-15, Trp-95, and Tyr-97 form a cluster, whose shape needs to be preserved for AraC to behave like WT.[14] Our work reported here and that reported in the literature suggest that the identification of such clusters can assist analysis of protein structure–function relationships.

Because restructuring of the front lip of the core does not seem to be a causative step in generating constitutivity, what is such a step? One way to look would be to examine the difference in the interaction energy matrices, for these can reveal changed interactions, including rearrangements primarily involving the side chains that may occur without large backbone alterations. Although hints of additional relevant interactions may be contained in the interaction matrices constructed for this work, because some allosteric interactions occur relatively slowly, such an approach might better be done with considerably longer simulation times than were possible in the current study.

In this work, using a combination of molecular dynamics, molecular genetics, and biochemical measurements, we conclude that the folded or unfolded state of the lip of the dimerization domain is not directly involved in determining whether the protein is in a repressing or inducing state. Examination of the crystal structures of the dimerization domain of AraC in the absence and presence of arabinose had previously suggested a similar conclusion.[10,11] It should be noted, however, that conclusions based on crystal structures are not entirely reliable. First, it is possible that crystallization trapped one or both of the dimerization domain structures in nonrepresentative conformations. That

is certainly the case for the first apo-AraC dimerization domain structure that was determined.[33] Second, although the structures of the core of the dimerization domain in the absence and presence of arabinose are highly similar (RMSD = 1.5 Å), they are not identical, and thus it is possible that the minor structural changes induced in the core of the dimerization domain in response to the binding of arabinose are what lead to induction. Overall, however, the restructuring of the core is not sufficient to produce constitutivity, as suggested by (1) the fact that the preponderance of constitutive mutations lie in the N-terminal arm, (2) the arm's structure in the crystal structures changes significantly plus and minus arabinose, and (3) the SGLD and experimental results described in this article showed that core mutants producing restructuring of core residues 37–42 were uninducible. All together, the available data lead to the hypothesis that the interaction of the N-terminal arm with something other than the front lip is the primary determinant of the inducing versus repressing state of AraC, but further study is required to establish the mechanism.

## ACKNOWLEDGMENTS

The authors thank Michael Rodgers for innumerable comments and suggestions, Benjamin T. Miller for maintaining the nodes, Mili Shah for suggestions on mathematical techniques, Bernard Brooks for enabling access to NIH computers, Peter Kutt for coding assistance, and students in Bob Schleif's lab for contributing to results on mutants.

## REFERENCES

1. Englesberg E, Irr J, Power J, Lee N. Positive control of enzyme synthesis by gene C in the L-arabinose system. J Bacteriol 1965;90:946–957.
2. Sheppard DE, Englesberg E. Further evidence for positive control of the L-arabinose system by gene *araC*. J Mol Biol 1967;25:443–454.
3. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 1961;3:318–356.
4. Dunn TM, Hahn S, Ogden S, Schleif RF. An operator at -280 base pairs that is required for repression of *araBAD* operon promoter: Addition of DNA helical turns between the operator and promoter cyclically hinders repression. Proc Natl Acad Sci USA 1984;81:5017–5020.
5. Dunn TM, Schleif R. Deletion analysis of the *Escherichia coli* $araP_C$ and $P_{BAD}$ promoters. J Mol Biol 1984;180:201–204.
6. Martin K, Huo L, Schleif RF. The DNA loop model for *ara* repression: AraC protein occupies the proposed loop sites *in vivo* and repression-negative mutations lie in these same sites. Proc Natl Acad Sci USA 1986;83:3654–3658.
7. Lobell RB, Schleif RF. DNA looping and unlooping by AraC protein. Science 1990;250:528–532.
8. Carra JH, Schleif RF. Variation of half-site organization and DNA looping by AraC protein. EMBO J 1993;12:35–44.
9. Seabold RR, Schleif RF. Apo-AraC actively seeks to loop. J Mol Biol 1998;278:529–538.
10. Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C. Structural basis for ligand-regulated oligomerization of AraC. Science 1997;276:421–425.
11. Weldon JE, Rodgers ME, Larkin C, Schleif RF. Structure and properties of a truely apo form of AraC dimerization domain. Proteins 2007;66:646–654.
12. Ross JJ, Gryczynski U, Schleif R. Mutational analysis of residue roles in AraC function. J Mol Biol 2003;328:85–93.
13. Dirla S, Chien JY, Schleif R. Constitutive mutations in the *Escherichia coli* AraC protein. J Bacteriol 2009;191:2668–2674.
14. Damjanovic A, Miller BT, Schleif R. Understanding the basis of a class of paradoxical mutations in AraC through simulations. Proteins 2013;81:490–498.
15. Wu X, Brooks BR. Self-guided Langevin dynamics simulation method. Chem Phys Lett 2003;381:512–518.
16. Wu X, Damjanovic A, Brooks BR. Efficient and unbiased sampling of biomolecular systems in the canonical ensemble: A review of self-guided Langevin dynamics. In: Rice S, Dinner A, editors. Advances in chemical physics, Vol. 50. John Wiley & Sons; 2012. pp 255–326.
17. Lee MS, Olson MA. Protein folding simulations combining self-guided Langevin dynamics and temperature-based replica exchange. J Chem Theory Comput 2010;6:2477–2487.
18. Damjanovic A, Wu X, García-Moreno EB, Brooks BR. Backbone relaxation coupled to the ionization of internal groups in proteins: A self-guided Langevin dynamics study. Biophys J 2008;95:4091–4101.
19. Damjanovic A, García-Moreno EB, Brooks BR. Self-guided Langevin dynamics study of regulatory interactions in NtrC. Proteins 2009;76:1007–1019.
20. Pendse PY, Brooks BR, Klauda JB. Probing the periplasmic-open state of lactose permease in response to sugar binding and proton translocation. J Mol Biol 2010;404:506–521.
21. Damjanovic A, Miller BT, Wenaus TJ, Maksimovic P, Garcia-Moreno EB, Brooks BR. Open science grid study of the coupling between conformation and water content in the interior of a protein. J Chem Inf Model 2008;48:2021–2029.
22. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The biomolecular simulation program. J Comput Chem 2009;30:1545–1614.
23. MacKerell AD, Bashford D, Dunbrack, RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wirkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
24. Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ERP. Solution conformation of wild-type *E. coli* Hsp70 (DnaK) chaperone complexed with ADP and substrate. Proc Natl Acad Sci USA 2009;106:8471–8476.
25. Rodgers ME, Schleif R. Solution structure of the DNA binding domain of AraC protein. Proteins 2009;77:202–208.
26. Hahn S, Dunn T, Schleif R. Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. J Mol Biol 1984;180:61–72.
27. Hahn S, Schleif R. *In vivo* regulation of the *Escherichia coli araC* promoter. J Bacteriol 1983;155:593–600.
28. Schleif RF, Wensink PC. Practical methods in molecular biology. New York: Springer-Verlag; 1981.
29. Kolodrubetz D, Schleif R. L-arabinose transport systems in *Escherichia coli* K-12. J Bacteriol 1981;148:472–479.
30. Rodgers ME, Schleif R. DNA tape measurements of AraC. Nucleic Acids Res 2008;36:404–410.
31. Saviola B, Seabold R, Schleif RF. Arm-domain interactions in AraC. J Mol Biol 1998;278:539–548.
32. Emrick MA, Lee T, Starkey PJ, Mumby MC, Resing KA, Ahn NG. The gatekeeper residue controls autoactivation of ERK2 via a pathway of intramolecular connectivity. Proc Natl Acad Sci USA 2006;103:18101–18106.
33. Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C. The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. J Mol Biol 1997;273:226–237.